(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－169172

(43)公開日　平成11年(1999)６月29日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｃ１２Ｎ　15/09		Ｃ１２Ｎ　15/00	Ａ
Ｇ０１Ｎ　33/50		Ｇ０１Ｎ　33/50	Ｐ

審査請求　未請求　請求項の数11　ＯＬ　（全 10 頁）

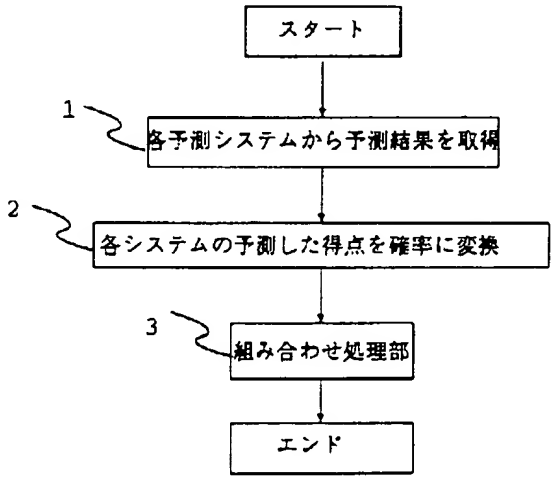
(21)出願番号　特願平9－336858

(22)出願日　平成９年(1997)12月８日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　村上　勝彦
東京都国分寺市東恋ヶ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　高木　利久
東京都港区白金台４－６－１
(74)代理人　弁理士　平木　祐輔

(54)【発明の名称】　ＤＮＡ塩基配列上のタンパク質コード領域予測方法及び記録媒体

(57)【要約】
【課題】　本発明の課題は、ＤＮＡ塩基配列上のタンパク質コード領域を正確に予測するコード領域予測方法を提供すること。
【解決手段】　異なるアルゴリズムを用いた複数のコード領域予測方法の結果を入力として１、あらためてコード領域を予測するコード領域予測方法であって、各予測方法から出力されるスコアを予測領域が正解である確率に変換し２、確率の値を比較する３ことで信頼度の高い領域を選択して予測する。これによって、各方法に用いられているアルゴリズムを組み合わせることが容易にでき、その結果正解率を上げることができる。



【特許請求の範囲】

【請求項1】　ＤＮＡ塩基配列データ上でタンパク質をコードする領域（これをコード領域と呼ぶ）を検出する方法において、異なるアルゴリズムを用いた複数のコード領域予測方法の予測結果を入力として、あらためてコード領域を予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項2】　請求項1に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各予測方法が共通にコード領域と予測した領域をとり、その領域をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項3】　請求項1に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各予測方法のいずれかが予測した領域のすべてをとり、その領域をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項4】　請求項1に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、あらかじめ各予測方法に優先順位をつけておき、もし、重なる領域を複数の予測方法がコード領域と予測した場合、その境界の決定については、優先順位の高い予測方法の予測による境界を採用してコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項5】　請求項1乃至4のいずれかに記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各コード領域予測方法がコード領域と予測した領域に与えられたスコアを、あらかじめ定めた関数によってその領域が正しくコード領域である確率に変換し、その確率の平均値があらかじめ定めた閾値よりも大きい時にその領域をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項6】　請求項1に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、もし、重なる領域を複数の予測方法がコード領域と予測した場合、その境界の決定については各コード領域予測方法がコード領域と予測した領域に与えられたスコアをあらかじめ定めた関数によってその領域が正しくコード領域である確率に変換し、確率が高い予測方法の予測による境界を採用してコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項7】　請求項6に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、その選択したコード領域予測方法の確率があらかじめ定めた閾値よりも大きい時にその領域をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項8】　請求項6又は7に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各コード領域予測方法がコード領域と予測した領域に与えたスコアを確率に変換する関数が、コード領域予測方法と、その方法が与えるスコアと、予測した領域の5'側及び3'側の境界の種類の関数であることを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項9】　請求項5、7又は8に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各予測方法の確率の平均値をその領域の最終スコアとするものであって、コード領域という予測を陽性、非コード領域という予測を陰性として、ＴＰを真陽性の塩基数、ＴＮを真陰性の塩基数、ＦＰを偽陽性の塩基数、ＦＮを偽陰性の塩基数、ＰＰを陽性の塩基数、ＰＮを陰性の塩基数、ＡＰをコード領域全体の塩基数、ＡＮを非コード領域全体の塩基数として、（ＴＰ)(ＴＮ)－(ＦＰ)(ＦＮ)を（ＰＰ)(ＰＮ)(ＡＰ)(ＡＮ）の平方根で割ったものとして定義されるファイ相関係数を計算した時に、この相関係数が既知の配列データを解析した場合に最大になるようなスコアの閾値をもうけておき、領域の最終スコアがこの閾値以上の場合に、その領域をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項10】　請求項5、7又は8に記載のＤＮＡ塩基配列上のタンパク質コード領域予測方法において、各予測方法の確率の平均値をその領域の最終スコアとして、ＦＰを偽陽性の塩基数、ＦＮを偽陰性の塩基数、ＡＰをコード領域の塩基数、ＡＮを非コード領域の塩基数とするとき、Ｅ＝（ＦＮ／ＡＰ＋ＦＰ／ＡＮ）／2によって定義される平均誤り率が既知のデータを解析した場合に最低になるような閾値をもうけ、領域の最終スコアがこの閾値以上をコード領域と予測することを特徴とするＤＮＡ塩基配列上のタンパク質コード領域予測方法。

【請求項11】　請求項1乃至10のいずれかに記載されたＤＮＡ塩基配列上のタンパク質コード領域予測方法をコンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はＤＮＡ塩基配列の情報処理、特に配列上のタンパク質コード領域を検出する方法に関する。

【0002】

【従来の技術】現在、ゲノム解析計画によってＤＮＡ配列が大量に決定されている。何の情報も付加されていないＤＮＡ塩基配列のデータに対して、その中でタンパク質がコードされている領域（コード領域）を予測することは、新しい遺伝子を発見し、医学、薬学的な研究をすすめる上で重要である。配列を決定した後で情報処理により遺伝子領域を推定できれば効率的に新しい遺伝子を特定することができる。ＤＮＡ配列中のコード領域を発見する方法としては、既知の核酸配列データベースに対して似た配列があるかどうかを検索する方法がある。新

しい配列の中で、既知の遺伝子に似ている配列は遺伝子である可能性が高いからである。しかし、長い領域にわたって問題のＤＮＡ配列と似ている遺伝子がデータベースにない場合にはこの方法は有効でない。近年、ＧＲＡＩＬ(Proceedings, The Second International Conference on Intelligent Systems for Molecular Biology, page376-384, 1994, ISBN 0-929280-68-7) という方法などのように、統計情報をもとにしてコード領域を見つける方法が発展してきた。

【０００３】このＧＲＡＩＬで使われるコード領域予測方法のアルゴリズムは、主にコード領域に現れる数塩基の配列の統計、すなわち６塩基の短い配列(以下、6merと呼ぶ) がコード領域に出現する頻度に基づいている。ある6mer (例えば、X=acgttc)がデータベース中の配列に出現した頻度のうち、コード領域に出て来た頻度Ｆｃ(Ｘ) と非コード領域に出て来た頻度Ｆｎ(Ｘ) の比Ｐ(Ｘ)=Ｆｃ(Ｘ)/Ｆｎ(Ｘ) をもって、この6merのスコアとする。さらに、6merでなく、ある程度長い領域がコード領域であるかどうかを識別するのには、その領域に現れた6merすべてについて、上記の6merのスコアを（数１）にあるように乗じて、この領域のスコアとする。

【０００４】

【数１】

$$CP_1=\prod_{i=0}^{k-2} P(a_{3i+1}\cdots a_{3i+5}a_{3i+6})P(a_{3i+2}\cdots a_{3i+6}a_{3i+7})P(a_{3i+3}\cdots a_{3i+7}a_{3i+8}) \quad \text{----------(数1)}$$

この数値をコーディングポテンシャルと言い、その値の大きさはその領域がコード領域らしい程度を示す。これ以外にもいくつか類似の指標を計算し、それらをフィードフォワード型のニューラルネットワークに入力することにより、最終的にその領域に０以上１以下の得点を与え、一定値以上であればコード領域であると予測する。

【０００５】他にも、同様にＤＮＡ配列からコード領域を予測する方法として、ＦＥＸＨ (Nucleic Acid Research, vol. 22, Num. 24, page 5156-5163, 1994)やGeneParser (Journal of Molecular Biology, 248, page 1-18, 1995)などがある。ＦＥＸＨで使われるコード領域予測方法のアルゴリズムでは、ＧＲＡＩＬとは異なるコーディングポテンシャルを用い、さらにコード領域の境界のシグナル検出方法を用いている。ＦＥＸＨでは8merのスコアを（数２）によって計算する。

【０００６】

【数２】

$$P(X)=\frac{F_c(X)}{F_c(X)+F_n(X)} \quad \text{----------(数2)}$$

ここで、Ｆｃ(Ｘ) は、8merＸがデータベース中のコード領域に出て来た頻度、Ｆｎ(Ｘ) は、8merＸがデータベース中の非コード領域に出て来た頻度である。ＦＥＸＨにおけるコーディングポテンシャルの計算は（数２）を用いて定義される。すなわち、任意の長さの配列について、その配列のコーディングポテンシャルは、（数２）を考える領域にわたって平均したもの（数３）である。

【０００７】

【数３】

$$CP_2=\frac{1}{N-7}\Sigma_{i=1}^{N-7}P(a_i,\ldots,a_{i+7}) \quad \text{----------(数3)}$$

ＦＥＸＨで使われるコード領域予測方法におけるコード領域の境界のシグナル検出方法について説明するが、その前に境界に関する基本的事柄を述べる。境界から5'側(左側)の塩基の位置を -1,-2,...と表現し、境界からすぐ3'側(右側)の塩基の位置を境界に近いほうから 1, 2,...と表現する。０の位置はない。以下では例えば、-4から+3までの塩基位置の７文字からなる領域は(-4,3)と表す。コード領域の境界では、境界からの位置によって使われる塩基の種類に偏りがあり、よく出て来る配列パターンを境界の「コンセンサス配列」という。コード領域の境界の種類はいくつかあって、原核生物の場合は開始コドンとよばれる配列'ＡＴＧ'で始まるか、終止コドンと呼ばれる配列で終わるが、ヒトを初めとする真核生物の場合はさらに、コード領域の5'側の境界と3'側の境界とで２種類あり、この２つはそれぞれアクセプターサイト，ドナーサイトと呼ばれる。この種類によって、コンセンサス配列が違う。ほとんどの場合、コード領域の間に存在する介在配列と呼ばれる領域は、ＧＴの２塩基で始まり、ＡＧの２塩基で終わるので、配列ＧＴが見つかればアクセプターサイトの候補であり、配列にＡＧが見つかればドナーサイトの候補である。これらのうち、実際にはコード領域の境界ではない位置を疑似境界部位と呼ぶ。

【０００８】ＦＥＸＨにおいては、コード領域のシグナルを検出するために、以下で定義される「3merのスコア」を用いる。まず、領域（Ｌ,Ｒ）に注目する。Ｌ,Ｒは領域の境界の位置である。ある3mer Y_k (k=1,...,64)が領域（Ｌ,Ｒ）内の位置ｉにでてくる頻度を、実際の境界のデータと疑似部位に対してそれぞれ $F^i_{s,k}$, $F^i_{p,k}$ とおき、「位置ｉでの3merのスコア」を（数４）

によって定義する。

【0009】

【数4】

$$P(i)=\frac{F^{i}_{s,k}}{F^{i}_{s,k}+F^{i}_{p,k}} \quad \cdots\cdots(数4)$$

この（数4）を領域（L,R）に渡って平均したもの（数5）が領域（L,R）の「3merのスコア」である。

【0010】

【数5】

$$P=\frac{1}{m}(\Sigma_{i=L}^{R}P(i)) \quad \cdots\cdots(数5)$$

ここで、mは領域（L,R）の長さである。さらに、（数5）の和をとるときに、ある値$\alpha=0.65$を与えておき、P(i)がαよりも大きいP(i)の和をとったときのスコアを特に、「有意な3merのスコア」と定義する。さて、FEXHで使われるコード領域予測方法におけるコード領域の境界のシグナル検出方法について説明する。ドナーサイトについては、(-30,-5)のコーディングポテンシャル、領域(-4,6)の3merのスコア、領域(-30,-5)の3merのスコア、領域(7,50)の3merのスコア、領域(-4,6)の有意な3merのスコア、領域(6,50)のG，GG，GGGの数などを計算し、統計的手法である判別分析によりこれらのスコアを組み合わせて、一つのドナーサイトに対して一つの統合スコアを計算する。アクセプターサイトについては、領域(-48,-34) の3merのスコア、領域(-33,-7)の3merのスコア、領域(-6,5)の3merのスコア、領域(6,30)の3merのスコア、領域(1,54)の8merのコーディングポテンシャル、領域(-1,-54)の8merのコーディングポテンシャル、領域(-33,-7)のT，Cの数を判別分析によって組み合わせて、一つのアクセプターサイトに対して一つの統合スコアを計算する。

【0011】ある領域がコード領域かどうかを決定するためのスコアは、コーディングポテンシャル、ドナーサイトのスコア、アクセプターサイトのスコアなどを判別分析で組み合わせて、統合スコアが閾値より高い領域をコード領域と予測する。GeneParser で使われるコード領域予測方法のアルゴリズムでは、i 塩基目からj塩基目までの領域のコーディングポテンシャルを計算するのに、（数6）を用いている。

【0012】

【数6】

$$IFhex(i,j)=\max\begin{cases}\Sigma_{k=0,3,\ldots,j-6}\ln(\frac{f_k}{F_k})\\ \Sigma_{k=1,4,\ldots,j-6}\ln(\frac{f_k}{F_k})\\ \Sigma_{k=2,5,\ldots,i-6}\ln(\frac{f_k}{F_k})\end{cases} \quad \cdots\cdots(数6)$$

ここでf_k は6mer $a_{i+k},a_{i+k+1},\ldots,a_{i+k+5}$ ($a_1\in\{A,C,G,T\}$）が学習データのコード領域に現れた頻度で、6merの第一番目の文字がコドンの一番目になっているときだけ数えたものである。F_k は、その配列と同じ塩基組成でランダムに配列を生成した場合にその6merが出て来る頻度の期待値である。

【0013】さらに、（数7）で定義される局所的複雑度も用いている。

【0014】

【数7】

$$H=-\sum_{k\in\{A,C,G,T\}}(\frac{N_k}{L})\log_2(\frac{N_k}{L}) \quad \cdots\cdots(数7)$$

ここで、Lは考慮している配列の長さで、N_kは塩基k（kは、A，C，G，Tのいずれか）が長さLの配列に現れた頻度である。GeneParser では、コード領域の境界のスコアを以下のように計算する。境界の領域(i, j) 内の配列$s_i s_{i+1}\ldots s_j$ に対してそのスコアS(i,j)は（数8）で定義する。

【0015】

【数8】

$$S(i,j)=\Sigma_{k=i}^{j}\log_2(f_{b,k}/p_b) \quad \cdots\cdots(数8)$$

ここで、$f_{b,i}$ は、位置iに塩基b（bはA，C，G，Tのいずれか）が出てきた頻度であり、p_b は塩基bがその配列に出てくる事前確率である。他にもコード領域らしいかどうかのスコアや、ドナーサイトのスコア、アクセプターサイトのスコアを計算し、それらをフィードフォワード型のニューラルネットワークで統合して、最終的にその領域に0以上1以下の得点を与え、一定値以上であればコード領域と予測結果を出す。GeneParserは、GRAILと同じくニューラルネットワークを用いているが、先に述べたように考慮している特徴が異なり、さらにその学習方法も異なっている。

【0016】このように、これらの予測方法は異なる特

徴検出方法を用いており、それらのスコアを統合化する方法も異なっている。これらの方法の予測は不正解であることも多く、同じ配列を解析しても、検出しにくいコード領域に対しては方法によって予測結果が異なることが多い。

【００１７】

【発明が解決しようとする課題】これまでは単一の方法をもとにしていたため、精度が低かった。これは、各予測方法の捉えている特徴が部分的なものだからである。したがって、できるだけ多くの特徴をとらえた結果をまとめて予測をして、正解率を上げることが課題である。一方、各方法で採用している配列の特徴検出方法を一つの方法の中で実装するのは、人的コストがかかる。そこで、出来るだけ簡便な方法で多くの特徴をとらえることが課題である。本発明の目的は、上記の課題を解決し、信頼性の高いコード領域予測をする方法を提供することが目的である。

【００１８】

【課題を解決するための手段】上記課題を解決するために本発明では、ＤＮＡ配列上にある遺伝子の様々な特徴を異なるアルゴリズムよって学習した遺伝子予測方法の結果を入力に用いて、それらを総合した予測を行う。これによって、一つの予測方法で考慮できない多くの特徴を考慮した予測結果を容易に得られる。また、異なる予測方法による予測結果の信頼性を比較したうえで最終予測を行うために、各予測方法のスコアを正解率に変換し、正解率を比較することによって、より信頼の高い予測結果を得る。

【００１９】

【発明の実施の形態】図１は本発明の実施の概要で、複数の予測プログラムの結果を使って、新たな予測をする方法の流れ図である。各予測方法による予測処理部１で、入力されたＤＮＡ配列データを各コード領域予測方法で解析し、各々の予測結果を得る。各コード領域予測プログラムの出力結果の中から、入力配列の何塩基目から何塩基目までがコード領域であるかという情報と、そのスコア(確からしさ) の情報を保持しておく。次に、スコア変換部２で、各予測された領域のスコアからその領域が正解である確率を求める。正解であるとは、予測した領域と実際のコード領域がオーバーラップしていることである。この確率は、あらかじめ設定してある変換関数で得られる。この確率は、以下Ｐscore と呼ぶ。変換関数の作成方法については、後述する。

【００２０】各予測結果をもとにして、あらためてコード領域を予測する組み合わせ処理３を行い、最終予測結果を出力して終了である。図２はこの組み合わせ処理３の詳細についての説明図である。図３は複数の予測方法によって、異なる領域がコード領域と予測された例である。図３の上部に、３つのコード領域予測方法によってコード領域と予測された領域を描いた。すなわち、ＦＥＸＨによってコード領域と予測された領域21、GeneParser によってコード領域と予測された領域22、ＧＲＡＩＬによってコード領域と予測された領域23である。それぞれのＰscore は、0.8 、 0.4、 0.9である。横軸は入力したＤＮＡ配列上での塩基の位置を示す。まず、記録配列24を用意し、その全領域に整数０を入れて初期化11を行う。次に、各予測方法で予測された各塩基に対応する記録配列24の部分に1 を足して予測結果を記録する処理12を行う。これによってどの場所がいくつの予測方法によってコード領域と予測されたかが、記録配列24に記録される。図３では、重なった予測方法の数を四角の高さで表しているが、実際には0 以上の整数が入っている。

【００２１】次に、スキャン処理13によって、記録配列24をスキャンし、１以上が記録されている重なり領域29をみつける。重なり領域29がなければ条件式14によって終了し、あればコード領域の境界を決定する処理15に進む。境界決定処理15と、最終スコアの決定16に関して、５つの方法を述べる。１つめの方法では、使用した全ての予測方法が予測した領域をコード領域の候補とする。このときの最終スコアは、各方法のＰscore の平均とする。この例では、最終スコアは(0.8+0.4+0.9)/3 = 0.7である。以下これを「ＡＮＤ法」と呼ぶ。この方法によってコード領域と予測され得る領域25を図３に示した。この時点では、まだ候補であって、コード領域と予測されたわけではない。２つめの方法では、使用した予測方法のいずれかが予測した領域をコード領域の候補とする。このときの最終スコアは、各方法のスコアから換算した各Ｐscore の平均とする。もし、いくつかの予測方法がその重なり領域中のどこもコード領域と予測していなかった場合の最終スコアは、予測しなかった予測方法のＰscore を０として、計算する。この例では、最終スコアは(0.8+0.4+0.9)/3 = 0.7 である。以下これを「ＯＲ法」と呼ぶ。この方法によってコード領域と予測され得る領域26を図３に示した。

【００２２】３つめの方法では、その重なり領域29をコード領域と予測した予測方法のうち、最も高いＰscore を持っている予測方法を選択し、その予測方法が予測した領域をコード領域の候補とする。このときの最終スコアは、ＯＲ法と同様に各Ｐscore の平均とする。もし、いくつかの予測方法がその重なり領域中のどこもコード領域と予測していなかった場合の最終スコアは、予測しなかった予測方法のＰscore を０として、計算する。この例では、最終スコアは(0.8+0.4+0.9)/3 = 0.7 である。以下これを「ＨＩＧＨＥＳＴ法」と呼ぶ。この方法によってコード領域と予測され得る領域27を図３に示した。

【００２３】４つめの方法では、その重なり領域30をコード領域と予測した予測方法のうち、あらかじめつけておいた優先順位の高い予測方法の結果を選択する。その

予測方法が予測した領域をコード領域の候補とする。優先順位は、あらかじめ各予測方法のテストをしておきその成績の良い順にする。例えば、バーセットらの行った遺伝子予測プログラムのテスト(Genomics, 34, page 353-367, 1996)において、境界も正確に予測したエキソン（コード領域の単位）の数の割合が高い順に設定すればよい。この順は、順位の高いほうからＦＥＸＨ，GeneParser，ＧＲＡＩＬである。このときの最終スコアは、選択した予測方法のPscore とする。この例では、最終スコアはＦＥＸＨのPscore になるので 0.8である。以下これを「ＲＵＬＥ法」と呼ぶ。この方法によってコード領域と予測され得る領域28を図３に示した。

【００２４】５つめの方法では、その重なり領域30それぞれに対して、Pscore と境界のタイプを考慮して新しいスコアEscore を計算する。境界のタイプとは、開始コドン、ドナーサイト、アクセプターサイト、終止コドンのうちのいずれかである。この４つを以下では、i,d,a,t と表記する。Escore は以下のように計算する。まず、ある予測方法が予測したコード領域のうち、コード領域のPscore がpsで、境界のタイプtypeにおける正解率をPe(type,ps)とする。これは、各予測方法毎に学習データから近似関数を作成しておく。この例では、ＦＥＸＨの境界のタイプがa,t でPscore が0.8 であるから、ＦＥＸＨのP(a,0.8) とP(t,0.8) を計算しておく。また、GeneParser の境界のタイプがa,d でPscore が0.4 であるからGeneParser のP(a,0.4) とP(d,0.4) を計算しておく。さらに、ＧＲＡＩＬの境界のタイプがi,t でPscore が0.9 であるから、ＧＲＡＩＬのP(i,0.9) とP(t,0.9) を計算しておく。次に、ある予測方法が予測した領域の左右の境界のタイプがl及びrならば（l,r はi,d,a,t のいずれかである）、このときのEscore は、（数９）と定義する。

【００２５】

【数９】

$$Escore = Pe(l, ps)Pe(r, ps) \quad \cdots\cdots(数9)$$

この例でＦＥＸＨ、GeneParser 、ＧＲＡＩＬのEscore は、それぞれ0.72,0.48,0.54である。このEscore が最も高い予測方法が予測した領域をコード領域の候補とする。このときの最終スコアは、選択した予測方法のEscore とする。この例では、最終スコアは0.8*0.9=0.72である。以下これを「ＥＤＧＥ法」と呼ぶ。この方法によってコード領域と予測される領域29を図３に示した。

【００２６】各方法のいずれかで候補の領域と最終スコアを決定した後、その候補をコード領域として予測するかどうかを閾値によって決定する。すなわち、最終スコアと閾値とを比較する部分17によって最終スコアが高いかどうかを判断し、高ければこれをコード領域と予測して、領域を出力する処理18を行う。ここで閾値の設定方法について述べる。閾値を設定するときに高く設定すると、偽陽性の数は減るが感度が悪くなる。逆に、閾値を低く設定すると、感度が上がるが偽陽性の数が増えてしまう。そこで、閾値は何らかの指標が最適になるような適当な値に決めなければならない。ここでは、多くの配列データに対して、予測と正解の相関を示すファイ相関係数の全データに渡っての平均が、最高になるように定める。このファイ相関係数は、正解の分かっている配列、すなわち、コード領域の位置が分かっているＤＮＡ配列一つに対して一つの値が求められる。また、このファイ相関係数の定義は、（数１０）である。

【００２７】

【数１０】

$$CC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}} \quad \cdots\cdots(数10)$$

ただし、コード領域という予測を陽性、非コード領域という予測を陰性として、ＴＰは真陽性の塩基数、ＴＮは真陰性の塩基数、ＦＰは偽陽性の塩基数、ＦＮは偽陰性の塩基数、ＰＰは陽性の塩基数、ＰＮは陰性の塩基数、ＡＰはコード領域の塩基数、ＡＮは非コード領域の塩基数とする。$-1 \leqq CC \leqq 1$であり、完全に正解であればＣＣ＝１であり、完全に正解と逆の予測をしたときはＣＣ＝－１となる。ランダムな予測に対しては、ファイ相関係数の期待値は０である。

【００２８】閾値の決定に際しては、各方法毎に、多くの配列データにおいてあらかじめ０から１までのいろいろな閾値でファイ相関係数の平均を求めて、ＣＣが最高値をとるときの閾値を、その方法の閾値と決める。閾値の決定に関しては、上記の方法の他に第１種の誤り率E_1＝ＦＮ／ＡＰと、第２種の誤り率E_2＝ＦＰ／ＡＮの平均E_{AV}＝（E_1＋E_2）／２　が最低値をとるときの閾値にする方法も考えられる。あるいは、上記の誤り率を計算する際に塩基数でなくコード領域の数で計算した誤り率の平均を最低値にするような閾値を採用する方法も考えられる。

【００２９】ここで、各予測方法のスコアを予測した領域が正解である確率に変換する関数の作成方法を図４に沿って述べる。多くのデータを解析する必要があるので、はじめにさまざまな条件を満たすデータのみを集める処理31を行う。例えば、核酸配列データベースであるGenBank リリース100 (1997 年4 月) の中から、項目'ＳＯＵＲＣＥ' が'Homo sapiens'であり、一つ以上の'ＣＤＳ' を含んでいるＤＮＡ配列のエントリーを集める。これらはコード領域の位置が分かっているＤＮＡ配列データである。また、イントロン領域を含むものについては、イントロン領域がＧＴで始まるかＡＧで終ることが条件であるので、この条件を満たさないデータは

捨てる。また、pseudo, putative, ＯＲＦ, alternative, predict,fusionのうちいずれかの記述が項目'ＣＤＳ'の中にあれば、それは実験的にコード領域とは確認されていないか、確実なコード領域がわかっていない可能性が強いため、これを除く。さらに、各コード領域予測方法が学習に用いたデータをここで使わないようにするため、1996年6月より前に登録されたデータを除く。

【００３０】これらの処理31を経てデータセット32を作成する。データ一つについて各予測方法での解析33を実行した後、各予測方法のスコアとエラー率の関係を調べる処理34をする。実際に調べたＦＥＸＨのヒストグラムを図５に、GeneParser のヒストグラムを図６に、ＧＲＡＩＬのヒストグラムを図７に示す。これらのヒストグラムから、各予測方法ごとに（数１１）で仮定した変換関数を求めるため最小自乗法によるパラメータ推定35を行う。こうして出来た関数が、求める変換関数である。この変換関数は、一次式でなく二次式でもよい。

【００３１】

【数１１】

$$Pscore(score)=A+B\times score \quad \cdots\cdots(数11)$$

本方法によって正解率がどの程度変わるかを示す。ＤＮＡ配列データセット32を本方法で解析し、配列一つごとに以下の正解率（数１２）を計算し、これをデータの本数で割ったものである。なお、この正解率の計算方法はバーセットらによって提案された正解率であり(Genomics, 34, page 353-367, 1996)、広く使われているものである。

【００３２】

【数１２】

$$AC=\frac{1}{2}\left[\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right]-1 \quad \cdots\cdots(数12)$$

各予測方法単独のときの数6 の値はＦＥＸＨが0.64、GeneParser が0.63、ＧＲＡＩＬが0.67であったのに対し、３方法の組合せを本発明のように行うと、ＡＮＤ法では 0.53 と低くなったが、ＯＲ法で 0.76 、ＨＩＧＨＥＳＴ法で 0.74 、ＲＵＬＥ法で 0.71 、ＥＤＧＥ法で0.74と高くなり、後者４つの方法では正解率が上がった。

【００３３】ＡＮＤ法は全体の正解率では、単独の結果より悪いが実際のコード領域のうち、検出できなかったコード領域の率をみると、単独では、ＦＥＸＨが0.47、GeneParser が0.45、ＧＲＡＩＬが0.27であるのに対し、ＡＮＤ法による３つのコード領域予測方法を組合せると、0.07と低くなり、93%のコード領域を検出することができた。ＡＮＤ方法は、感度と特定度のバランスでは、単独のときの正解率におよばないが、特別な場合、すなわち偽陽性が多くてもコード領域として可能性のある領域を出来るだけ多くリストアップしたい場合には有効である。

【００３４】なお、上記以外にも他のアルゴリズムを用いたコード領域予測方法が知られているので、（表１）に例示する。これらの予測方法も任意に選択して用いることができる。

【００３５】

【表１】

（表１）他のアルゴリズムを用いたコード領域予測方法

名前	方法	開発元
ＥＲ	線形判別分析	東大（日本）
ＧＥＮＳＣＡＮ	マルコフモデル	スタンフォード大学
GenLang	確定節文法	ペンシルベニア大学
GenView	線形結合	イタリア国立研究会議（伊）
GeneＩＤ	パーセプトロン	ボストン大学
GeneMark	マルコフモデル	ジョージア工科大学
Genie	隠れマルコフモデル	カリフォルニア大学
ＭＯＲＧＡＮ	決定木	ジョンズホプキンス大学
ＭＺＥＦ	２次判別分析	コールドスプリングハーバー研究所
ＯＣ１	決定木	ジョンズホプキンス大学
PROCRUSTES	スプライスアライメント	南カリフォルニア大学
Sorfind	ルールベース	ラビットハッチ・コーポレーション
ＶＥＩＬ	隠れマルコフモデル	ジョンズホプキンス大学

【００３６】

【発明の効果】従来、一つのアルゴリズムによる場合は少ない指標にもとづいて予測していたため、効率良く多くのコード領域を得ることができなかった。本発明によ

って、複数のアルゴリズムで多角的に候補を選定できるので偽陽性があまり増えずに効率的に多くのコード領域を検出できる。その結果、全体的に精度が上がる。

【図面の簡単な説明】

【図１】本発明におけるコード領域予測手順

【図２】組合せ処理の説明図

【図３】重なり領域での境界決定方法の説明図

【図４】スコアを確率へ変換する関数を作成する方法の説明図

【図５】予測方法ＦＥＸＨのスコアとエラー率の関係を示す図である。

【図６】予測方法GeneParser のスコアとエラー率の関係を示す図である。

【図７】予測方法ＧＲＡＩＬのスコアとエラー率の関係を示す図である。

【符号の説明】

1 …ＤＮＡ配列を各予測方法で解析する処理部

2 …予測方法のスコアを正解である確率に変換する部分

3 …組み合わせ処理部

11…記録配列の初期化

12…各方法の予測領域を記録配列に記録する処理

13…記録配列のスキャン処理

14…重なり領域のチェック

15…境界の決定

16…最終スコアの決定

17…最終スコアと閾値の比較

18…コード領域と判定した領域の出力

21…予測方法ＦＥＸＨが予測したコード領域

22…予測方法GeneParser が予測したコード領域

23…予測方法ＧＲＡＩＬが予測したコード領域

24…ＤＮＡ配列に対応する記録配列

25…「ＡＮＤ法」によってコード領域と予測される領域

26…「ＯＲ法」によってコード領域と予測される領域

27…「ＨＩＧＨＥＳＴ法」によってコード領域と予測される領域

28…「ＲＵＬＥ法」によってコード領域と予測される領域

29…「ＥＤＧＥ法」によってコード領域と予測される領域

30…各予測方法による予測の重なり領域

31…核酸データベースから条件にあうデータを取得する処理

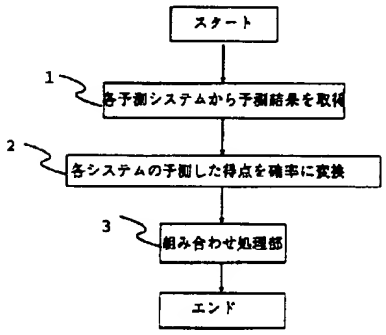
32…条件にあうデータの集合

33…各予測方法で解析する部分

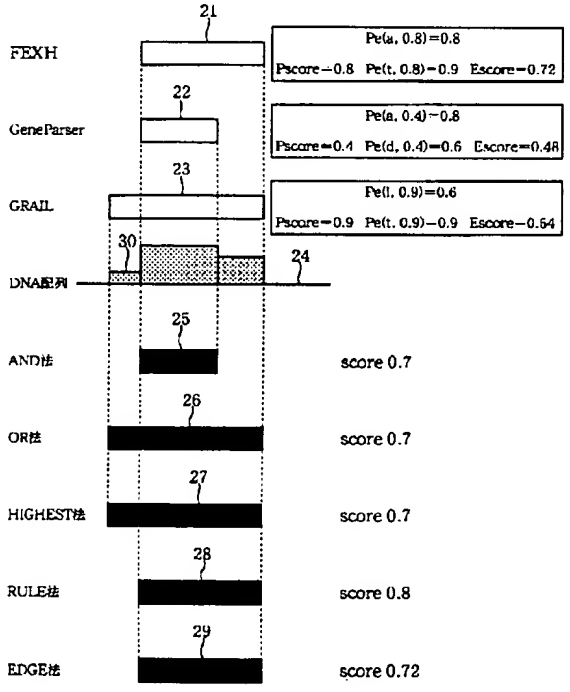
34…エラー率とスコアの関係を計算する部分

35…変換関数のパラメータを計算する部分

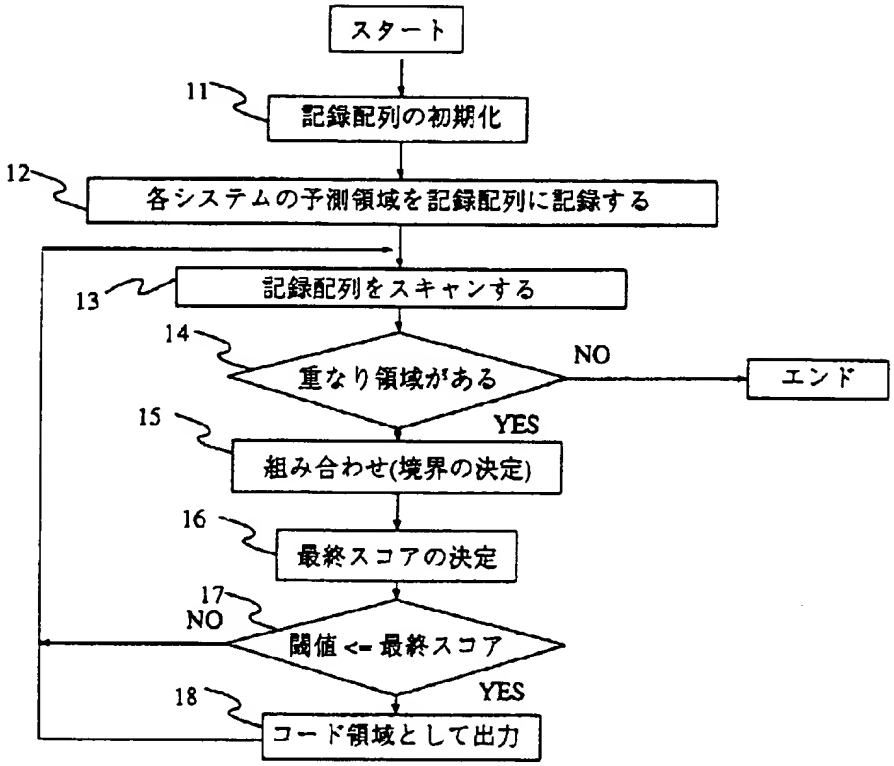
【図１】



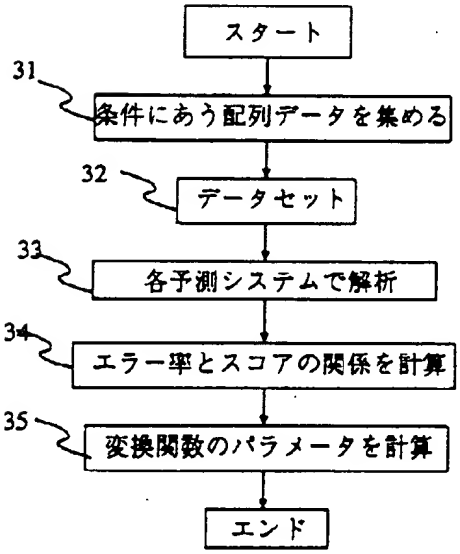
【図３】



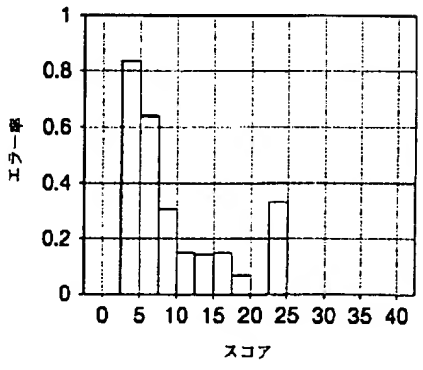
【図２】


スタート
11
記録配列の初期化
12
各システムの予測領域を記録配列に記録する
13
記録配列をスキャンする
14
重なり領域がある
NO
エンド
YES
15
組み合わせ(境界の決定)
16
最終スコアの決定
17
NO
閾値 <= 最終スコア
YES
18
コード領域として出力

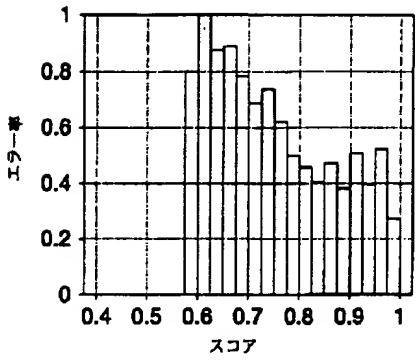
【図４】


スタート
31
条件にあう配列データを集める
32
データセット
33
各予測システムで解析
34
エラー率とスコアの関係を計算
35
変換関数のパラメータを計算
エンド

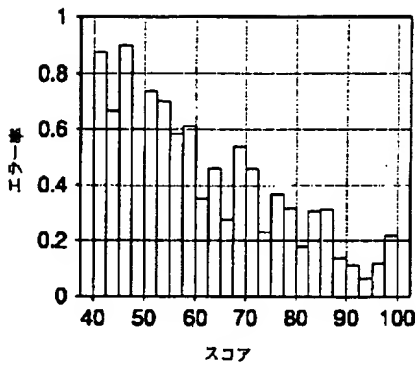
【図５】


1
0.8
0.6
0.4
0.2
0
エラー率
0 5 10 15 20 25 30 35 40
スコア

【図6】


1
0.8
0.6
0.4
0.2
0
エラー率
0.4 0.5 0.6 0.7 0.8 0.9 1
スコア

【図7】


1
0.8
0.6
0.4
0.2
0
エラー率
40 50 60 70 80 90 100
スコア